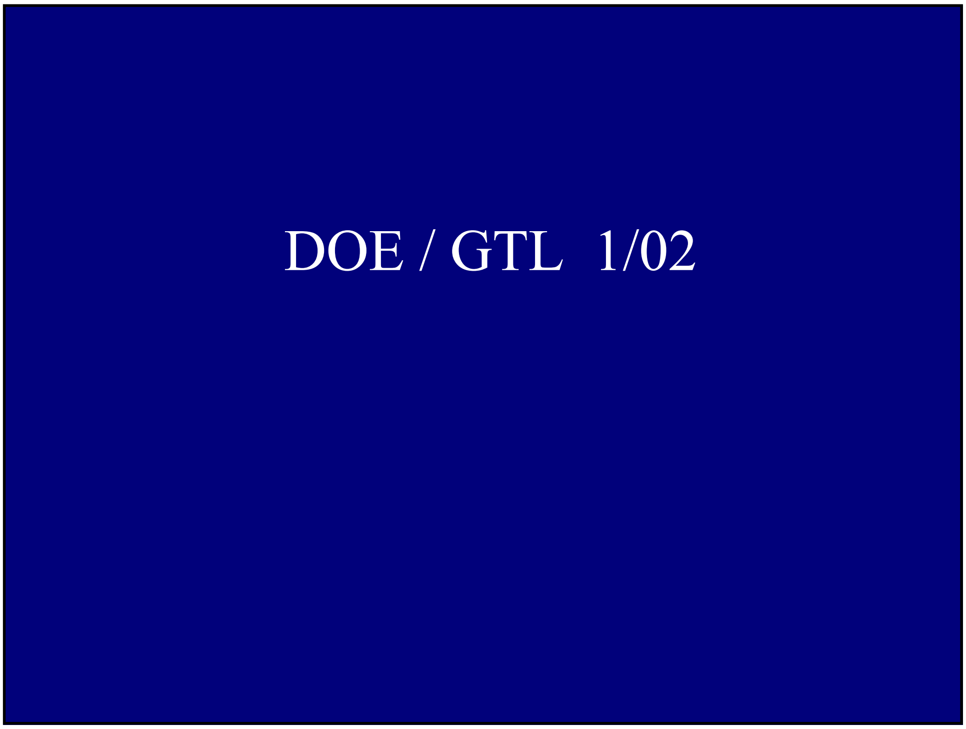
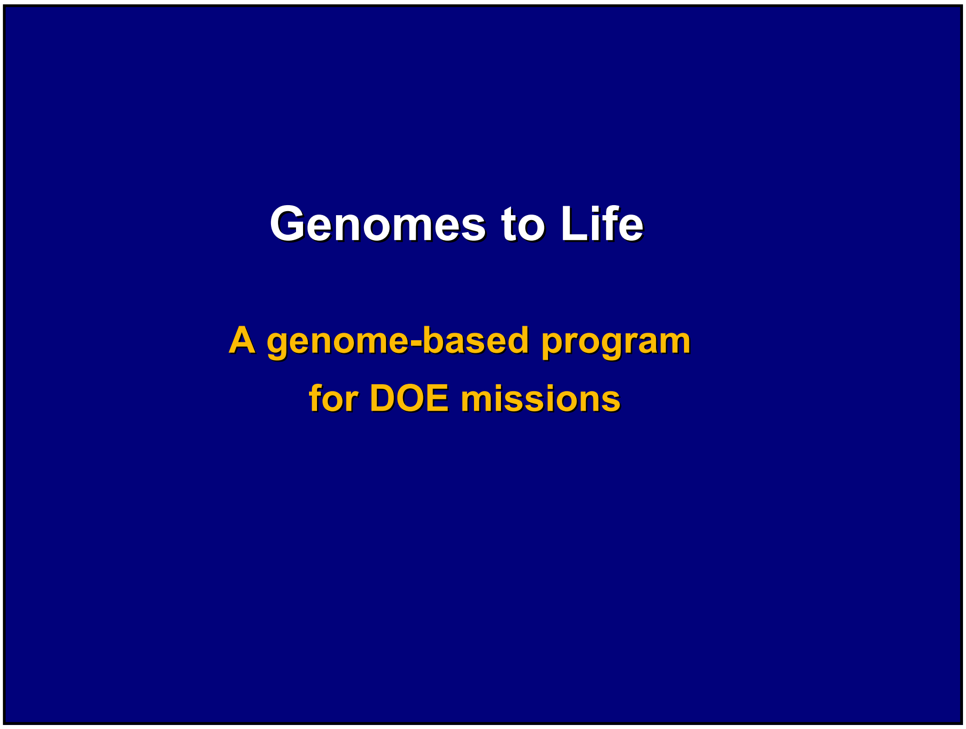
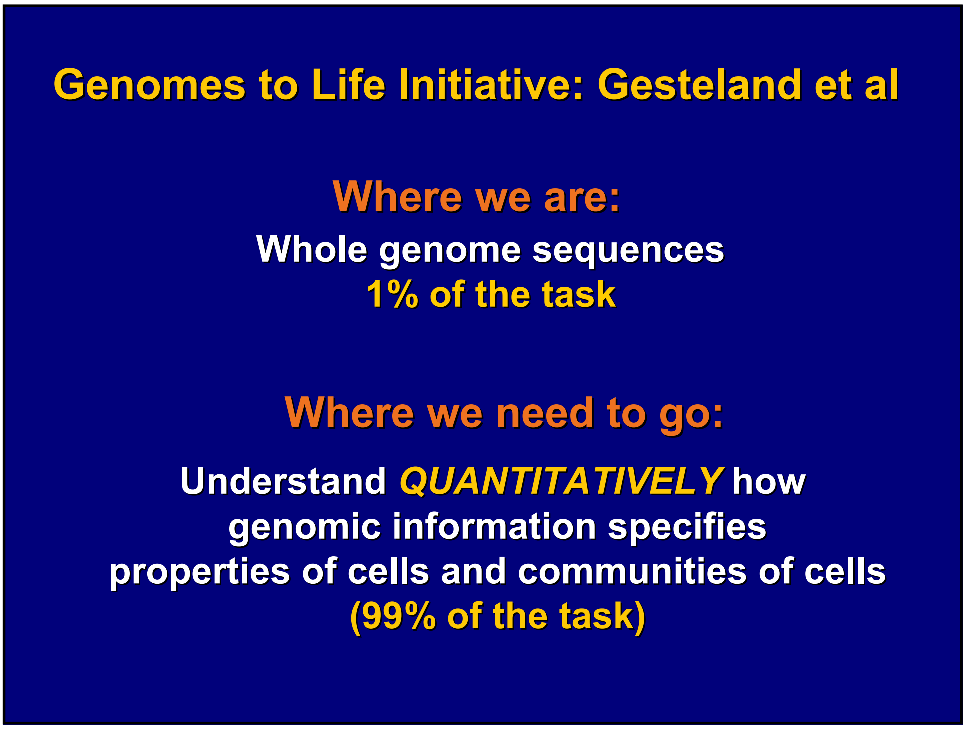


# DOE / GTL 1/02



# Genomes to Life

A genome-based program
for DOE missions



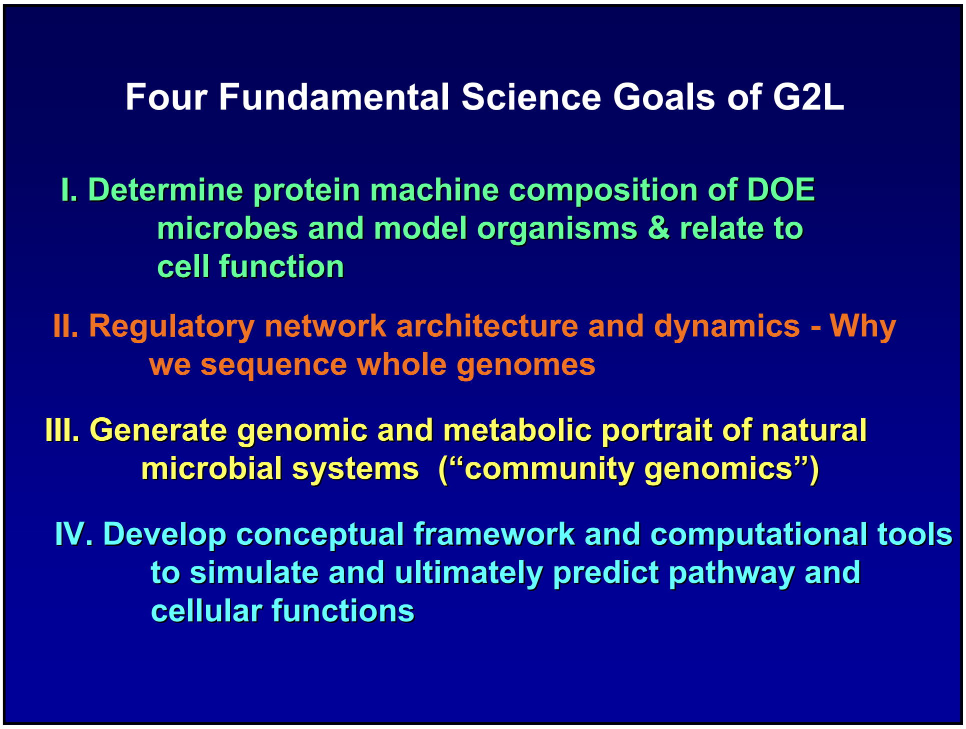
## Genomes to Life Initiative: Gesteland et al

### Where we are:

Whole genome sequences
1% of the task

### Where we need to go:

Understand *QUANTITATIVELY* how
genomic information specifies
properties of cells and communities of cells
(99% of the task)



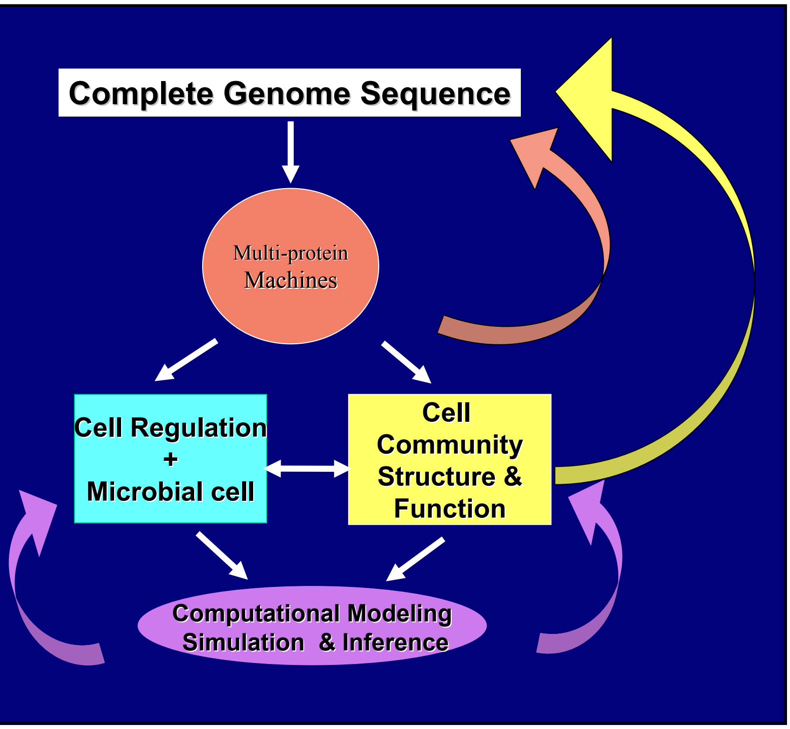
## Four Fundamental Science Goals of G2L

I. Determine protein machine composition of DOE microbes and model organisms & relate to cell function

II. Regulatory network architecture and dynamics - Why we sequence whole genomes

III. Generate genomic and metabolic portrait of natural microbial systems (“community genomics”)

IV. Develop conceptual framework and computational tools to simulate and ultimately predict pathway and cellular functions


Complete Genome Sequence

Multi-protein Machines

Cell Regulation + Microbial cell

Cell Community Structure & Function

Computational Modeling Simulation & Inference

## Bringing “Genomes to Life” to life

$20MY to start

Lab call out

University call out this week

Up to 2/3 to each call in review

$1-6MY pilot centers

## Goal 1 Biology Premises:

1. Most proteins work as part of multi-protein complexes

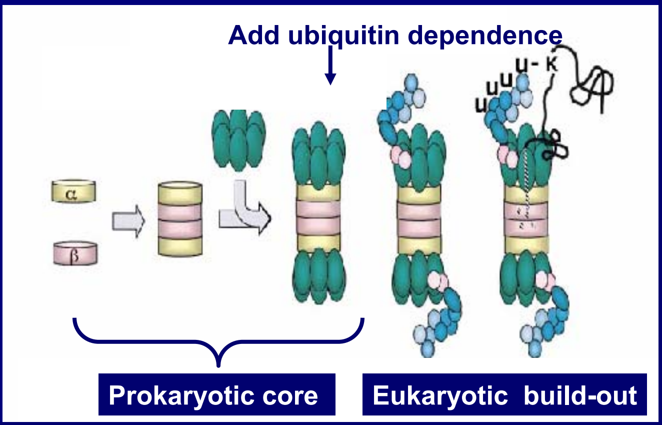
   Comprehensive knowledge of these
   is fundamental to understanding any cell

2. Number of <u>types</u> of machines believed finite

3. Significant core set of complexes are similar across evolution

## Goal 1 - Protein Machines

Well known case study: the Proteasome



Add ubiquitin dependence

Prokaryotic core

Eukaryotic build-out

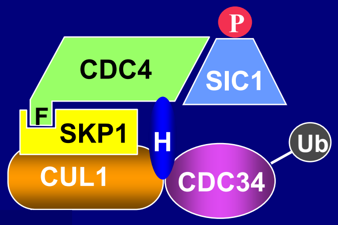
Function =protein garbage disposal of bacteria, plants, anima

## Map entire proteomes for multiprotein complexes

Theme protein complex + many variations

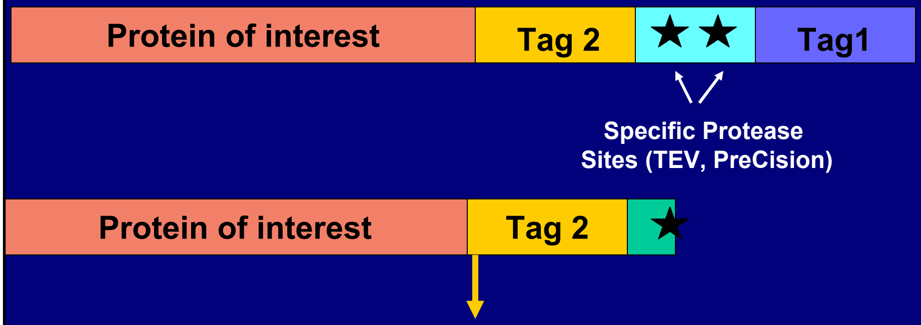
Example: SCF complexes (target specific proteins to proteasome□)

Variable composition  in some components, others constant



- Which parts constant,  which “variable”?
- What is effect on function:  Generic AND Specific functions
- Dynamic nature of machine composition

## Tandem Affinity Mass Spectrometry (B. Seraphin & colleagues)

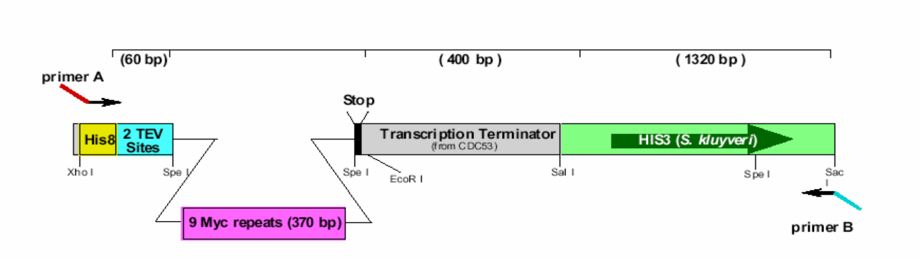


Specific Protease
Sites (TEV, PreCision)

Tag 25 : 20 mitosis; 5 transcription regulation

Tandem mass spec with prior chromatographic (J. Yates)
Separation - computational deconvolution of protein IDs

## Dual affinity tag version 1



Version 2 tests PreCision proetease / altered numbers of HIS and MYC epitopes

↓

Tag 25 : 20 for mitosis function;
5 transcription regulation
Several in versions with both tags

---

**ADA Chromatin Remodeling Complex**

| #21 Gcn5 (m) | %covered |
|---|---|
| YGR252W/GCN5 | 45.30% |
| YDR176W/NGG1 | 39.20% |
| *YDR448W/ADA2* | 32.00% |
| *YCL010C* | 16.60% |
| YOR023C/AHC1 | 14.50% |
| ? YBR066C/NRG2 | 8.60% |

**This study - 300 ml Conventional Mass spec**

**90 liters**

**SAGA Complex + others**

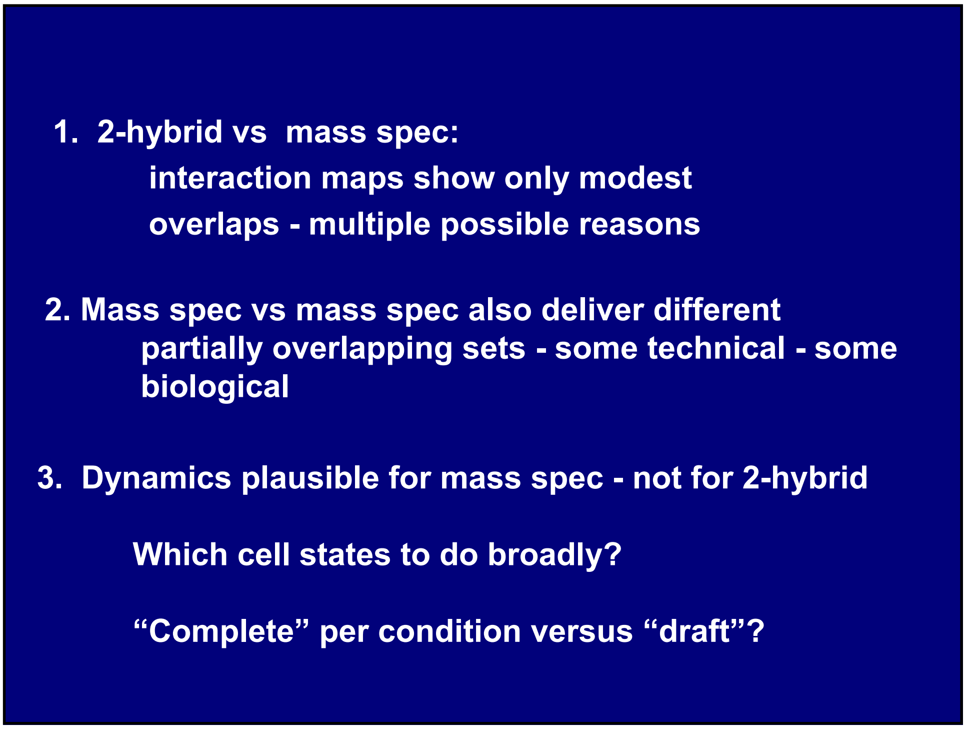
General issue of multiple complexes
that share some components -

Reciprocal tagging should help
to resolve some of this

## All corroborated interactions are among “certain” calls “uncertain calls” = blue: none had independent validation

| #20 Ynl116 (m) | %covered | #21 Gcn5 (m) | %covered |
|---|---|---|---|
| **YNL116W/ XXX** | 54.20% | **YGR252W/GCN5** | 45.30% |
| YHR115C/ XXX | 25.50% | **YDR176W/NGG1** | 39.20% |
| ***YNL311C/ XXX*** | 17.70% | ***YDR448W/ADA2*** | 32.00% |
| YML003W | 13.10% | ***YCL010C*** | 16.60% |
| **YDR328C/SKP1** | 10.30% | **YOR023C/AHC1** | 14.50% |
| YMR013C/SEC59 | 7.90% | YDR432W/ NPL3 | 13.00% |
| YGL180W/APG1 | 5.80% | YJR072C/XXX | 9.60% |
| YAL002W/VPS8 | 3.40% | YBR066C/NRG2 | 8.60% |
| YGR274C/TAF145 | 3.20% | YCL004W/PGS1 | 6.30% |
| YLR419W/XXXa | 2.80% | YPR112C/MRD1 | 6.00% |
| | | YBR017C/KAP104 | 3.90% |
| | | YDL102W/CDC2 | 3.60% |
| | | YPL074W/YTA6 | 3.20% |
| | | YKR099W/ BAS1 | 1.70% |
| | | YPR024W/YME1 | 1.50% |

| #3 Cdc20 (m) | %covere | #9 Glc7 (m) | %covere | #17 Sds22 (m,c) | %covere | #10 Ipl1 (no) | %covere | #15 Pds1 (m) |
|---|---|---|---|---|---|---|---|---|
| **YJL030W/MAD2** | 29.10% | ***YKL193C/ SDS22g*** | 99.10% | **YKL193C/ SDS22** | 96.70% | | | **YDR113C/PDS1** |
| **YGL116W/ CDC20o** | 26.40% | **YER133W/ GLC7g** | 93.90% | ***YER133W/ GLC7*** | 95.20% | | | YNL121C/TOM70 |
| ***YIL142W/CCT2*** | 21.30% | ***YOR227W*** | 69.00% | ***YFR003C/ XXX*** | 64.50% | | | YDR104C/SPO71 |
| ***YJL013C/MAD3*** | 11.30% | YER177W/BMH1 | 67.40% | YML016C/ PPZ1 | 51.30% | | | **YGR098C/ESP1** |
| ***YDR212W/TCP1*** | 6.60% | ***YJL042W/MHP1*** | 60.90% | YPL179W/PPQ1 | 35.50% | | | YNR031C/ SSK2 |
| YDL143W/CCT4 | 5.90% | ***YMR311C/GLC8*** | 59.00% | ***YDR436W/ PPZ2*** | 22.40% | | | YHR020W/ XXX |
| YJL008C/CCT8 | 3.70% | YDR099W/BMH2 | 55.70% | YJR119C | 8.00% | | | |
| | | YDR475C | 52.30% | YJL052W/TDH1 | 6.90% | | | SRP1(m) |
| CCT3 (m) | | YPL137C/XXXg | 49.50% | YHR037W/PUT2 | 5.00% | | | |
| CCT5 (m) | | YGR237C | 48.20% | YJR152W/ DAL5o | 5.00% | | | |
| MDH1(m) | | YDR474C | 47.40% | YEL060C/PRB1 | 4.90% | | | |
| MKK(m) | | **YDR195W/REF2** | 46.20% | YBR259W | 4.10% | | | |
| CCT4,7,8 filtered | | YKL018W | 45.00% | YOR317W/FAA1 | 3.40% | | | |
| | | ***YFR003C/ XXXg*** | 41.30% | YIL129C/ TAO3 | 2.40% | | | |
| | | **YNL233W/BNI4** | 41.00% | YHR020W/ XXXkm | 2.00% | | | |
| | | YGR156W/PTI1 | 38.80% | YIL091C | 1.90% | | | |
| | | YIL154C/IMP2 | 38.40% | YOR086C | 1.80% | | | |
| | | ***YDR028C/REG1*** | 28.20% | | | | | |
| | | YBL092WRPL32 | 27.70% | fyv14 | | | | |
| | | YNL178W/RPS3 | 26.20% | hxt6(m) | | | | |
| | | **YAL043C/PTA1** | 24.20% | Net1(m) | | | | |
| | | YNL222W/SSU72 | 19.40% | NSR1(m) | | | | |
| | | **YKL059C** | 18.80% | PMA1(m) | | | | |
| | | YER158C | 18.70% | PMA2(m) | | | | |
| | | YAL031C/FUN21 | 16.70% | REG1(m) | | | | |
| | | **YER054C/GIP2** | 13.10% | RSE1(m) | | | | |
| | | YLR075W/RPL10 | 12.20% | RVB1(m) | | | | |
| | | **YIL045W/PIG2** | 9.30% | SNF4(m) | | | | |
| | | **YLR277C/YSH1** | 7.30% | YGR130(m) | | | | |
| | | **YKR002W/PAP1** | 6.90% | YHR186(m) | | | | |
| | | YML010W/ SPT5g | 6.50% | | | | | |
| | | YAL035W/ FUN12k | 6.00% | | | | | |
| | | **YLR115W/CFT2** | 5.20% | | | | | |
| | | YML016C/ PPZ1g | 5.20% | | | | | |
| | | YBR073W/RDH54 | 4.90% | | | | | |
| | | YLR384C/IKI3 | 3.70% | | | | | |
| | | ***YLR430W/SEN1*** | 2.60% | | | | | |



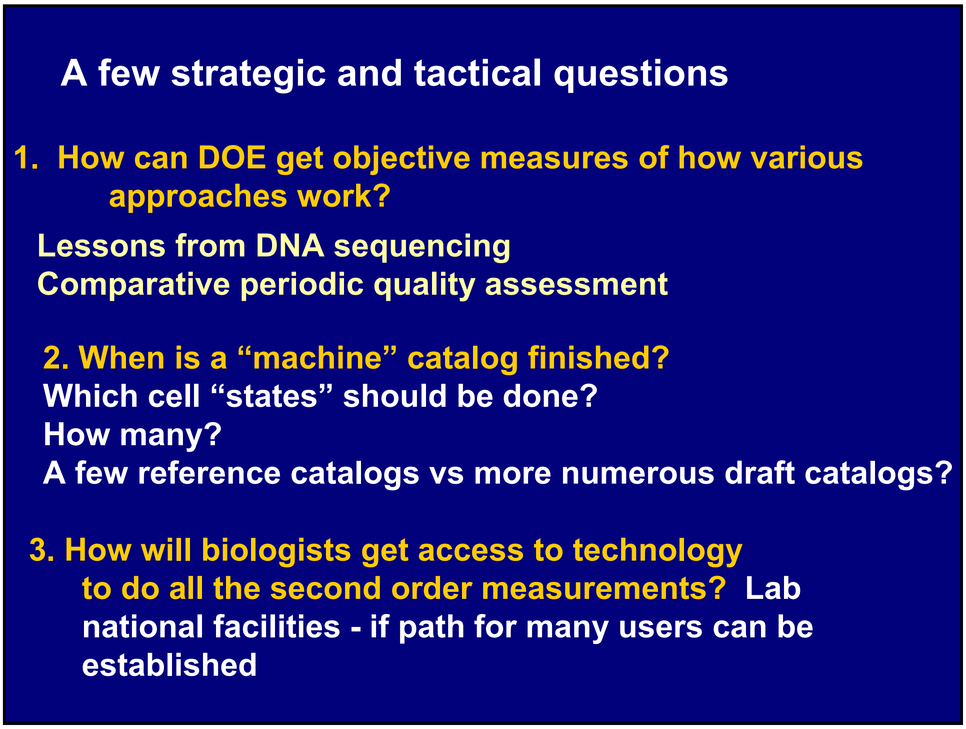
1. 2-hybrid vs mass spec:
   interaction maps show only modest overlaps - multiple possible reasons

2. Mass spec vs mass spec also deliver different partially overlapping sets - some technical - some biological

3. Dynamics plausible for mass spec - not for 2-hybrid

   Which cell states to do broadly?

   “Complete” per condition versus “draft”?



## A few strategic and tactical questions

1. How can DOE get objective measures of how various approaches work?

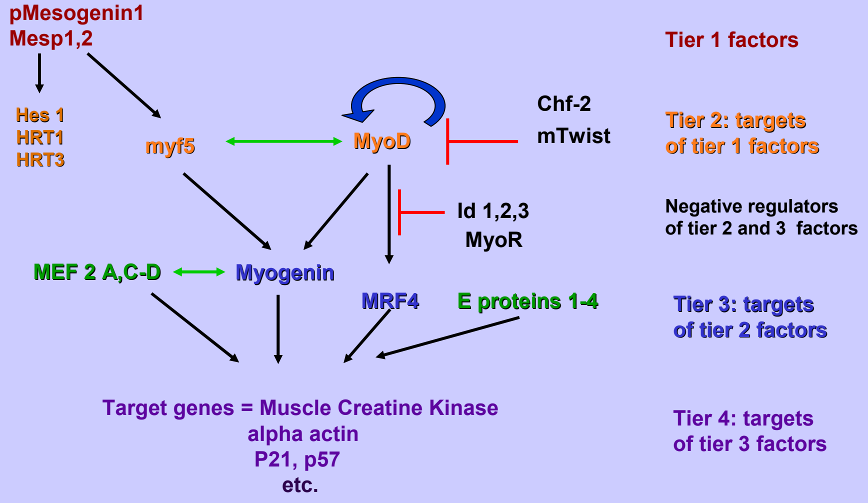
Lessons from DNA sequencing
Comparative periodic quality assessment

2. When is a “machine” catalog finished?
Which cell “states” should be done?
How many?
A few reference catalogs vs more numerous draft catalogs?

3. How will biologists get access to technology to do all the second order measurements? Lab national facilities - if path for many users can be established

## Goal 2. Regulatory network architecture and dynamics -

## B-HLH and MADS myogenic regulatory network



pMesogenin1
Mesp1,2

Tier 1 factors

Hes 1
HRT1
HRT3

myf5

MyoD

Chf-2
mTwist

Tier 2: targets
of tier 1 factors

Id 1,2,3
MyoR

Negative regulators
of tier 2 and 3 factors

MEF 2 A,C-D

Myogenin

MRF4

E proteins 1-4

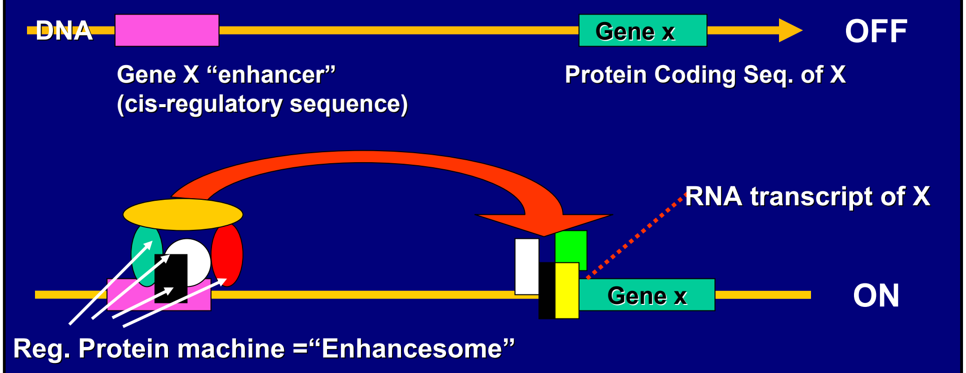
Tier 3: targets
of tier 2 factors

Target genes = Muscle Creatine Kinase
alpha actin
P21, p57
etc.

Tier 4: targets
of tier 3 factors

## Genes as complex informational entities

### Capturing Protein:DNA and Protein:Protein interactions that regulate activity



DNA

Gene x

OFF

Gene X "enhancer"
(cis-regulatory sequence)

Protein Coding Seq. of X

RNA transcript of X

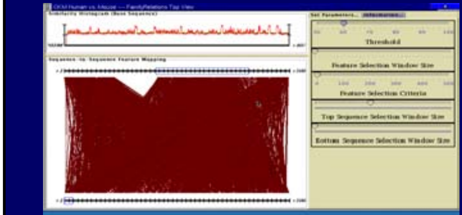
Gene x

ON

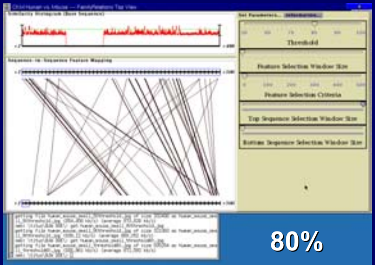
Reg. Protein machine ="Enhancesome"

Mouse/ Human pair-wise resolution at 60, 70, 80, 90% similarity in 50bp sliding feature over 50kb by 50kb region -

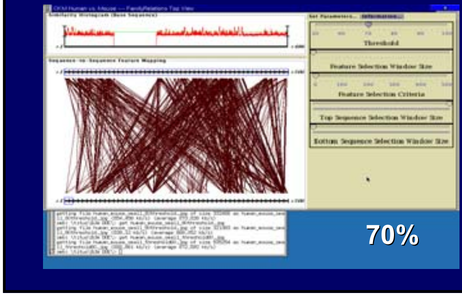
8 exons, proximal and distal cis-regulatory sequences


Similarity Histogram (Base Sequence)
Sequence-to-Sequence Feature Mapping
Threshold
Feature Selection Window Size
Feature Selection Criteria
Top Sequence Selection Window Size
Bottom Sequence Selection Window Size

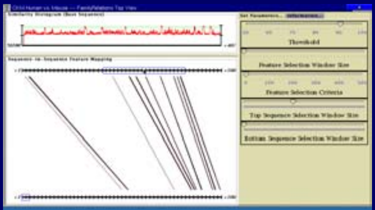
60% threshold


Similarity Histogram (Base Sequence)
Sequence-to-Sequence Feature Mapping
Threshold
Feature Selection Window Size
Feature Selection Criteria
Top Sequence Selection Window Size
Bottom Sequence Selection Window Size

80%

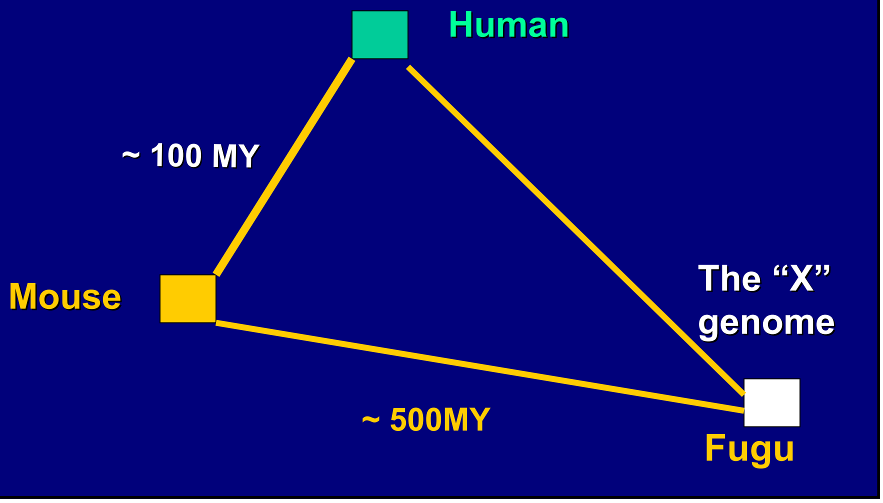

Similarity Histogram (Base Sequence)
Sequence-to-Sequence Feature Mapping
Threshold
Feature Selection Window Size
Feature Selection Criteria
Top Sequence Selection Window Size
Bottom Sequence Selection Window Size

70%


Similarity Histogram (Base Sequence)
Sequence-to-Sequence Feature Mapping
Threshold
Feature Selection Window Size
Feature Selection Criteria
Top Sequence Selection Window Size
Bottom Sequence Selection Window Size

90%

## Comparative genomics applied to finding cis-regulatory elements



2 Pairwise Similarity Maps

Simultaneous Triple-filter
@ 70% threshold

Fugu

Mouse

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

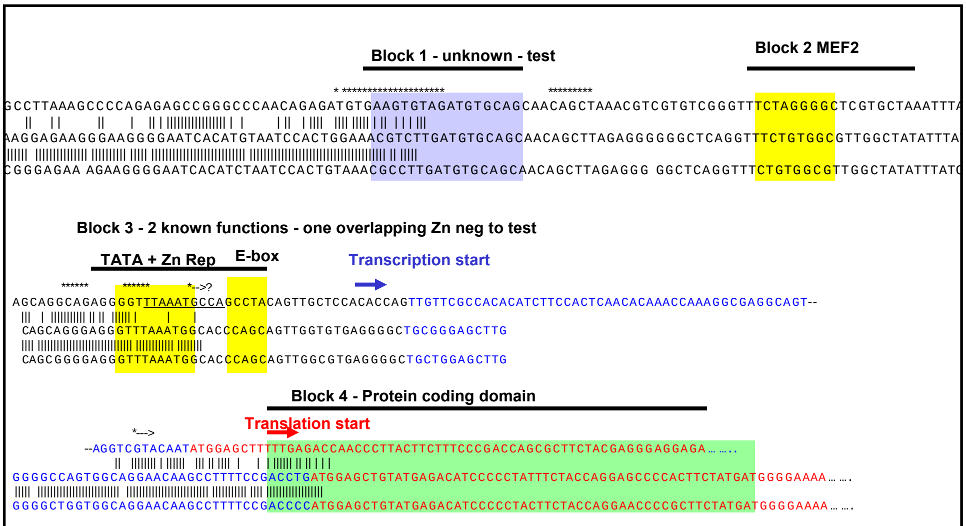
Human

FamilyRelations interactive comparisons
Titus Brown
Tristan DeByusscher

Sequence level inspection

## Close-up sequence inspection for myogenin

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.



Block 1 - unknown - test

Block 2 MEF2

```
                                               * *********************                     *********
GCCTTAAAGCCCCAGAGAGCCGGGCCCAACAGAGATGTGAAGTGTAGATGTGCAGCAACAGCTAAACGTCGTGTCGGGTTTCTAGGGGCTCGTGCTAAATTTA
   ||   | |      ||     |   || | |||||||||||||||| | |       | ||   | ||||   |||| ||||| | ||  | | | |||
AAGGAGAAGGGAAGGGGAATCACATGTAATCCACTGGAAACGTCTTGATGTGCAGCAACAGCTTAGAGGGGGGCTCAGGTTTCTGTGGCGTTGGCTATATTTA
||||||  |||||||||||||| |||||||| ||||| ||||||||||||||||||||||||| |||||||||||||||||||||||||||| || |||||
CGGGAGAA AGAAGGGGAATCACATCTAATCCACTGTAAACGCCTTGATGTGCAGCAACAGCTTAGAGGG GGCTCAGGTTTCTGTGGCGTTGGCTATATTTATC
```

Block 3 - 2 known functions - one overlapping Zn neg to test

TATA + Zn Rep E-box

Transcription start

```
      ******        ******       *-->?
AGCAGGCAGAGGGGTTTAAATGCCAGCCTACAGTTGCTCCACACCAGTTGTTCGCCACACATCTTCCACTCAACACAAACCAAAGGCGAGGCAGT--
 |||  | |||||||||| || ||  |||||| |      |     |
 CAGCAGGGAGGGTTTAAATGGCACCCAGCAGTTGGTGTGAGGGGCTGCGGGAGCTTG
 |||| ||||||||||||||||||||||||||||| |||||||||| ||||||||
 CAGCGGGGAGGGTTTAAATGGCACCCAGCAGTTGGCGTGAGGGGCTGCTGGAGCTTG
```

Block 4 - Protein coding domain

Translation start

```
              *--->
          --AGGTCGTACAATATGGAGCTTTTTGAGACCAACCCTTACTTCTTTCCCGACCAGCGCTTCTACGAGGGAGGAGA… …..
              ||  |||||||| | ||||||   ||| || ||||  |    |  | |||||| || || | | |
GGGGCCAGTGGCAGGAACAAGCCTTTTCCGACCTGATGGAGCTGTATGAGACATCCCCCTATTTCTACCAGGAGCCCCACTTCTATGATGGGGAAAA… ….
|||||  |||||||||||||||||||||  ||||||||||||||||||||||| ||||||||||| |||| ||||||||||||||||
GGGGCTGGTGGCAGGAACAAGCCTTTTCCGACCCCATGGAGCTGTATGAGACATCCCCCTACTTCTACCAGGAACCCCGCTTCTATGATGGGGAAAA… ….
```

**A third genome is powerful sieve for shared elements**

**Different regions of genome have changed at different rates:**
**implications for automation**

**Pufferfish is at outer limits of “too far” for cis-elements over noise,**
**but in some genes can be very effective**

## Lentiviral mediated mouse transgenesis ala Lois et al Science (online)

7/8 embryos
Positive

QuickTime™ and a
Photo - JPEG decompressor
are needed to see this picture.

## Low efficiency expression for conserved element on its own

Two copies of conserved element

Drive expression in somites - but
Unevenly compared with parent
Enhancer/promoter element

# GDF8 /  Myostatin

(Sejin Lee group - Johns Hopkins)

TGF beta family

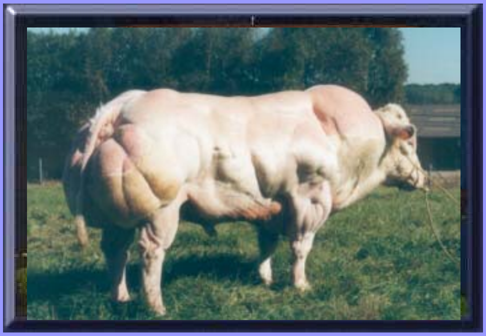
Paracrine factor

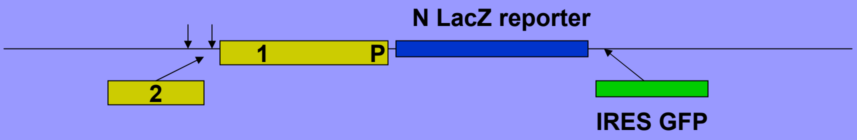
Activin receptor →

Less is more null phenotype

Barnyard natural variation

Gorilla? Bonobo? Olympian?





QuickTime™ and a
Photo - JPEG decompressor
are needed to see this picture.

## How many genomes at what distances do we need?

Collaboration with Paul Sternberg, Hiroke Shyzuya

Immediate goal Added Nematode genomes -
Large insert library resources for lateral comparisons of five genomes.

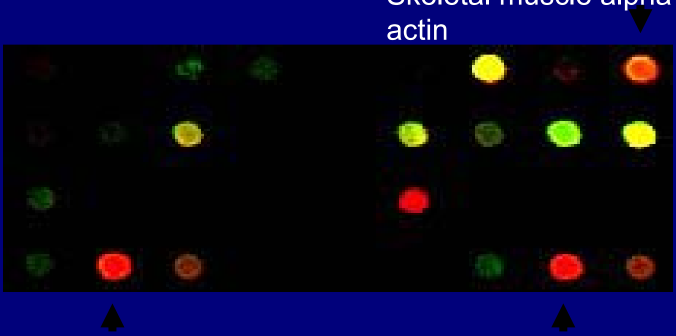
QuickTime™ and a
Photo - JPEG decompressor
are needed to see this picture.

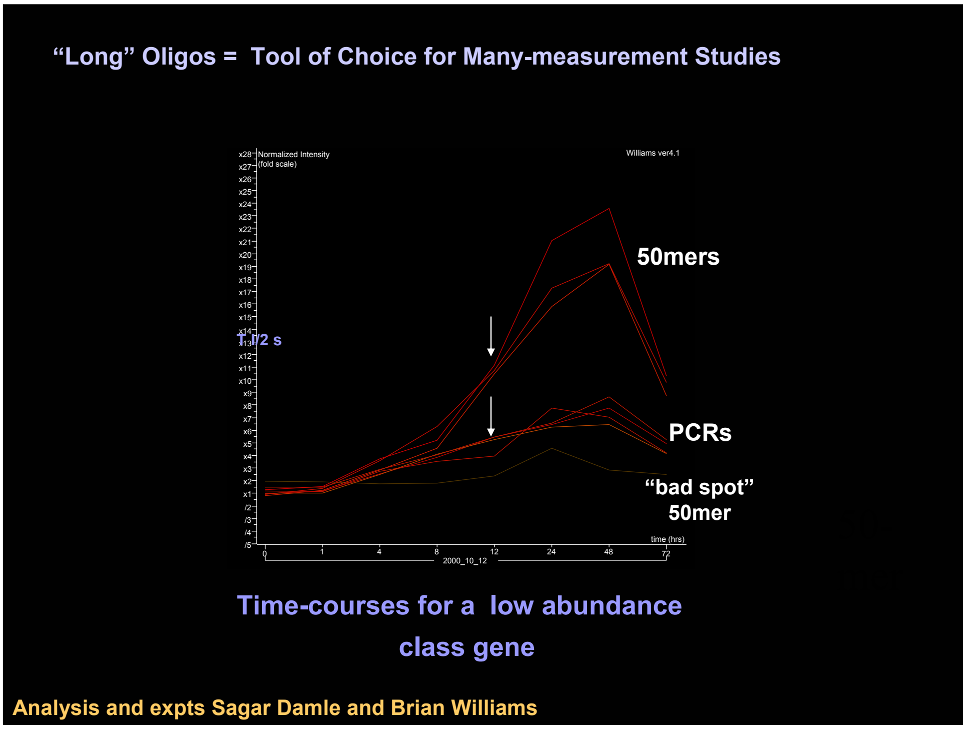
PS 1010 Fosmid library 15X coverage
positive screens for 3 test genes

CB5161 Fosmid library 11X coverage
Positive screens for 3 test genes

## Microarrays: current technology issue

Make case that for next 18 months, at least, “long” oligos are a superior strategy

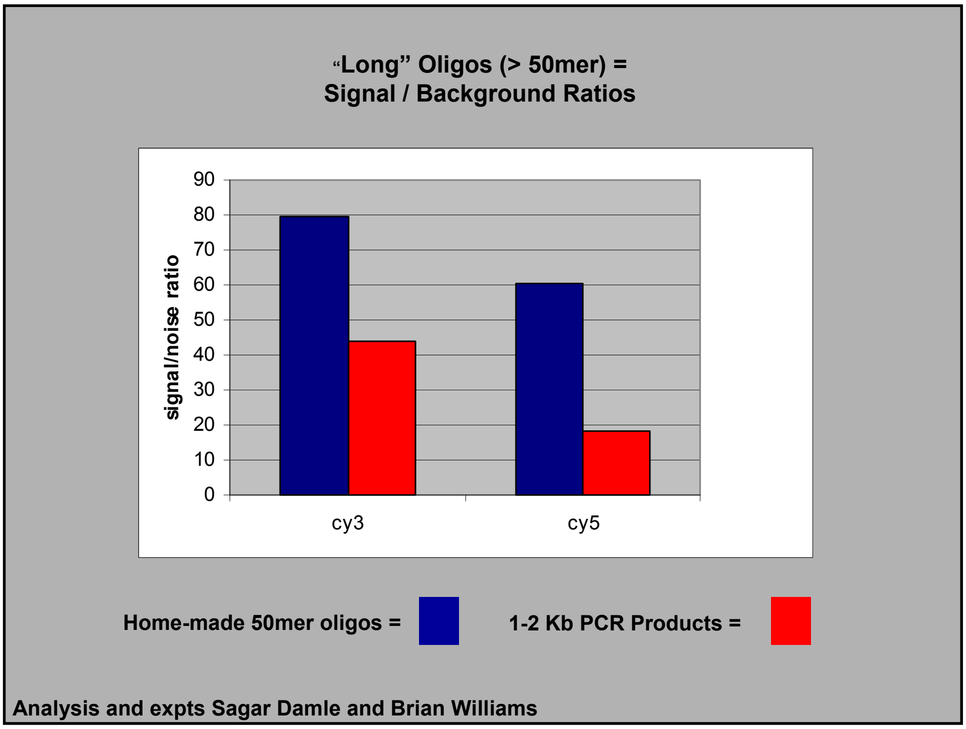




## “Long” Oligos = Tool of Choice for Many-measurement Studies

Time-courses for a low abundance class gene

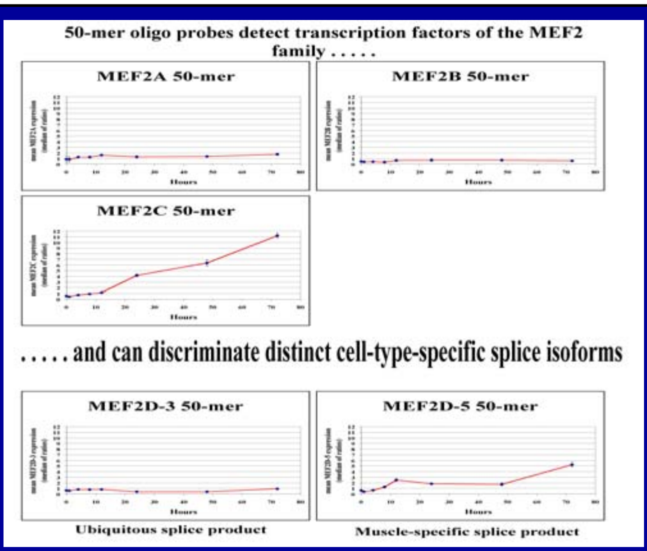
Analysis and expts Sagar Damle and Brian Williams



## “Long” Oligos (> 50mer) =
## Signal / Background Ratios

Home-made 50mer oligos = 🟦 1-2 Kb PCR Products = 🟥

Analysis and expts Sagar Damle and Brian Williams

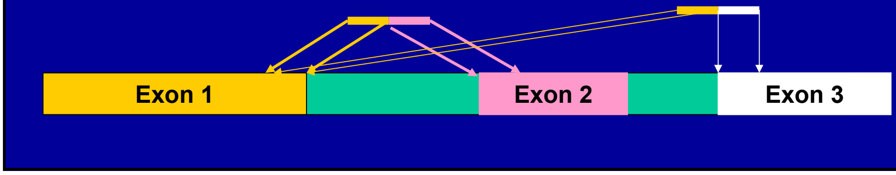


50-mer oligo probes detect transcription factors of the MEF2 family . . . . .

. . . . . and can discriminate distinct cell-type-specific splice isoforms

Ubiquitous splice product

Muscle-specific splice product



## “long” 50-70mer oligos currently a good strategy

1. Marginal cost per slide ~10X better than affymetrix

2. Marginal cost per slide ~ 4X better than PCR
   (plus reliability / repreducibility issues)

3. Options for splice isoform analysis superior

4. Option for specificty in gene families superior to PCR

5. Design option superior - two days from candidate
   sequence to array with new feature - no cloning
   intermediate

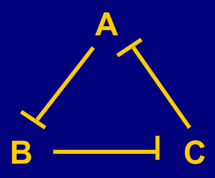
6. Technically superior to short (25mer) oligos
   because of specificity issues

**Still requires ratiometric (two color) measurements**

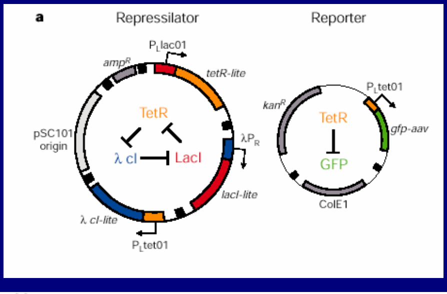
# Goal 4

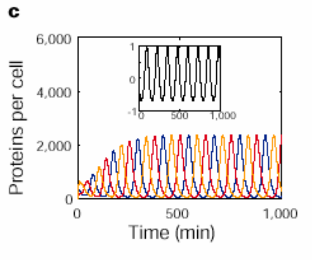
Develop conceptual framework and computational tools to simulate and ultimately predict pathway and cell functions

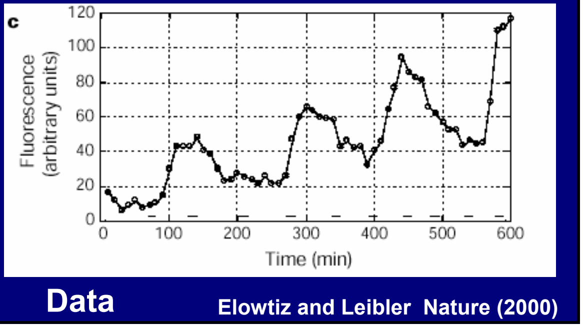
## Converting arrows and blockers to Computational predictive models of dynamic behavior



Oscillator







Theory

Data

Elowtiz and Leibler Nature (2000)

## The Age of Genes - 4 part PBS series

Peter Baker of Seeing Science media group

1. The public may need exposure to the questions more than the "answers"

2. "Context" can be "pre-considering" a problem by identifying with someone else's dilemma

3. Preventing misinformation and disinformation - partnership with FACs (Foundation for American Communications) - educate the journalists

Poster 184

Comparative genomics  software
Tristan Debuysscher   Triple view
Eric Davidson Group
Titus Brown   FamilyRel
Experimental:  Tristan,  Libera Berghella,  Tony Kirilusha

Microarray analysis

Brian Williams, Libera Berghella

Expression analysis, circuit modeling
Eric Mjolsness JPL group

Chris Hart
Ben Bornstein
Tobias Mann (now U. Wash)
Sagar Damale
Joe Roden
Becky Castano
Diane Trout

Mass spec analysis of protein complexes
Ray Deshiaes (Caltech)
Jea Hong Seoul
Leslie Dunipace
Johannes
John Yates (Scripps)
Hayes Mcdonald

Challenges in metagenomics of prokaryotes share
much with genomics of  - uneven representation of
Cell types that interact with each other in complex ways that
Are difficult to caputre in monoculture

## Scientific “opportunity space”

A. Whole Genome Sequences Available

B. Genome based biology - Now ready for Need Computation / Simulation

C. Massively parallel, high through-put technologies

## Why DOE for the goals of this program?

1. DOE congressionally mandated biological missions
2. Experience (climate; high thoughput biology)
3. Manpower (in labs, in academic collaborations)
4. Hardware (what DOE has now, future inventions)

Super Bac vector - szybalski arabinose inducible copy number
-get the vector and use for worm - can conjugate into subtlius -
Useful for Diane

Bacs 80K and over big enough to capture metabolic pathways
tend to be clustered eough to move whole trait: functional screening

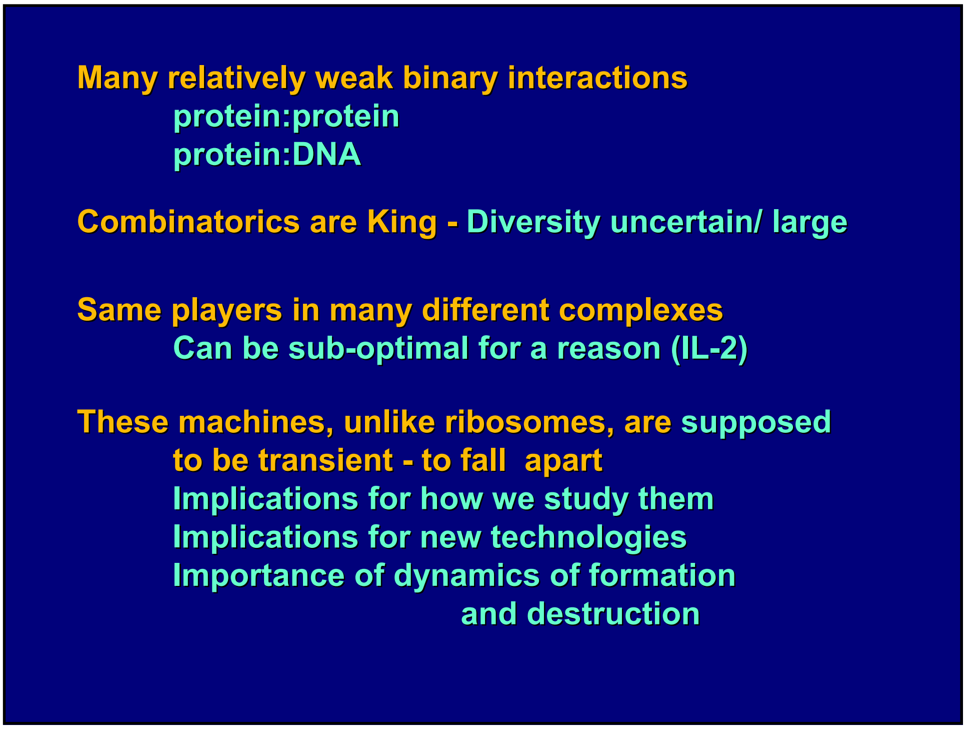
## Relationship of Goal 1 to other proteomics

**Whole cell proteomics to tell us what is there:**

**Measure in many cell states (microbial cell project)**

**e.g. biofilms versus dispersed cultures**

**status in communities versus monoculture**



Many relatively weak binary interactions
protein:protein
protein:DNA

Combinatorics are King - Diversity uncertain/ large

Same players in many different complexes
Can be sub-optimal for a reason (IL-2)

These machines, unlike ribosomes, are supposed
to be transient - to fall apart
Implications for how we study them
Implications for new technologies
Importance of dynamics of formation
and destruction